OLAP Cube Visualization of Hydrologic Data Catalogs

Ilya Zaslavsky[a], Matthew Rodriguez[a], Bora Beran[b], David Valentine[a], Jillian Wallis[c], and Catharine van Ingen[b]

San Diego Supercomputer Center, UCSD, San Diego, CA[a]; Microsoft Research, San Francisco, CA[b], and
Center for Embedded Networked Sensing, UCLA, Los Angeles, CA[C]

Abstract

As part of the CUAHSI Hydrologic Information System project, we assemble comprehensive observations data catalogs that support CUAHSI data discovery services and online mapping interfaces (e.g. the Data Access System for Hydrology, DASH). These catalogs describe several nation-wide data repositories that are important for hydrologists, including USGS NWIS and EPA STORET data collections. The catalogs contain a wealth of information reflecting the entire history and geography of hydrologic observations in the US. Enabling simple interactive data discovery across these catalogs is an integral step to making the data easily accessible for analysis.

OLAP (Online Analytical Processing), often referred to as data cube analysis, is an approach to organizing and querying large multi-dimensional data collections. We have applied the OLAP techniques, as implemented in Microsoft SQL Server 2005, to the analysis of the catalogs from several agencies. In this initial report, we focus on the OLAP technology as applied to catalogs, and preliminary results of the analysis.

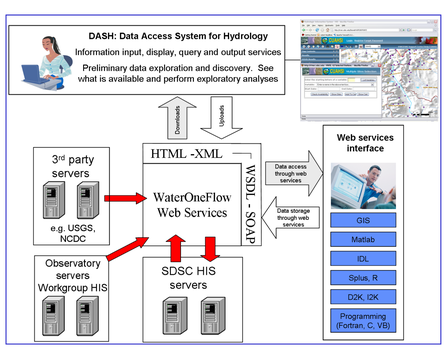
The initial results are related to hydrologic data availability from the observations data catalogs. The results reflect geography and history of available data totals from USGS NWIS and EPA STORET repositories, and spatial and temporal dynamics of available measurements for several key nutrient-related parameters.

About CUAHSI HIS

The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) is an organization representing 120+ universities in the US and abroad. As part of its mission, CUAHSI supports the development of cyberinfrastructure for the hydrologic sciences. The CUAHSI HIS (Hydrologic Information System) project is a multi-year multi-institution effort focused on consistent management of observations data available from several federal agencies (USGS, EPA, USDA, NOAA, etc.) as well as published by individual investigators.

CUAHSI HIS develops service-oriented architecture for hydrologic research and education, to enable publication, discovery, retrieval, analysis and integration of hydrologic data. The project team has defined a common information model for organizing hydrologic observation data, designed a common exchange protocol (Water Markup Language) and developed a collection of SOAP web services (WaterOneFlow services) that provide uniform access to different federal, state and local hydrologic data repositories.

This system is now implemented as a collection of Hydrologic Information Servers deployed at NSF-supported Hydrologic Observatory test beds.



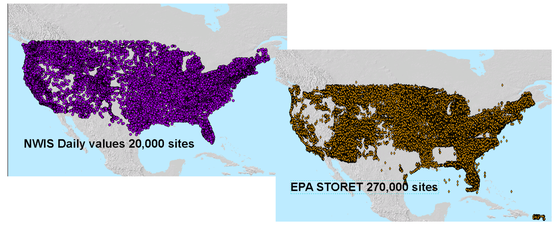
CUAHSI HIS Service Oriented Architecture: General Outline

OLAP Technology for Catalogs

Observations data catalogs assembled within CUAHSI HIS follow standard relational schema of the *Observations Data Model* (ODM). The ODM catalogs collections of observation series, each described by:

- *what:* observed parameter variable including metadata such as processing methodology, units, quality assessments
- *where:* location or site where the observation was made including relevant site metadata such as Hydrologic Unit or vegetation class
- *when:* start and end date of measurements and time granularity of measurement

OLAP (Online Analytical Processing) is an approach to rapidly analyze such large multi-dimensional databases. OLAP *Datacubes* organize data along *dimensions*, enable drilldown on *hierarchies*, and *aggregates* such as average or minimum. Using *SQL Server Analysis Services,* an OLAP cube can be created from any database with a star schema such as ODM. using *SQL Server Analysis Services*. Simple aggregates (such as daily averages or yearly maxima) can be pre-computed for improved query performance.



Using Datacubes

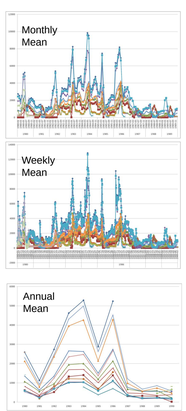
- Datacubes can be easily browsed with Excel PivotTables and drag-drop over the internet
- MatLab and other rich analysis client access integration coming
- Browsing can give fast overview insight into data availability and quality

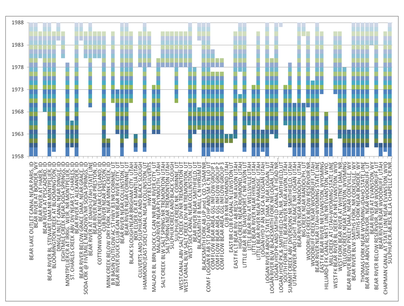
One of the things that a datacube enables is simple temporal drilldown. These examples are for the ODM sample database from the Little Bear River region in Utah (2 sites, 12 series, 593,000 observations).

The three graphs to the right plot mean annual, monthly, and weekly discharge at the dozen highest flow gages in the watershed over the years 1980 to 1990. The annual graph clearly shows the wet years - years of highest flow. The monthly and weekly graph show differences between years. For example, the high flows in 1984 occur in the middle of the year as the result of snow melt. The high flows in 1986 begin in the early part of the year before the usual snow melt continue through the summer season.

Daily Discharge Averaged by Year, Month, Week 1980-1990 (Top 12 Sites Only)

Drilling down time hierarchy gives overall sense of wet years and flood events





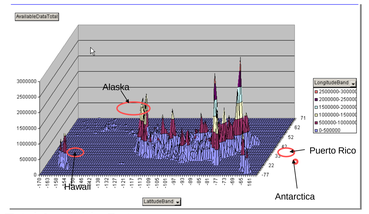
Bear River Sample Watershed stream gage comings and goings 1958-1988

To the left, data availability for a parameter, discharge, for a site shows that data availability is not always continuous. Sites have disappeared, and reappeared.

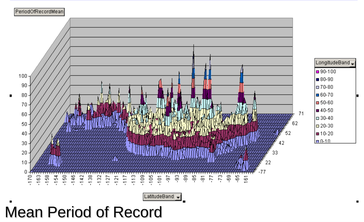
USGS NWIS Daily Values:

Mapping Data Availability

US Map of USGS Daily Values Series 1 degree latitude-longitude bins



Total Available Data



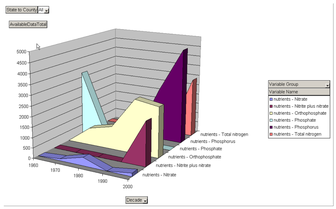
Mean Period of Record

These are examples from the USGS National Water Information System datacube. The cube contains 1.4 million sites and 12 million series.

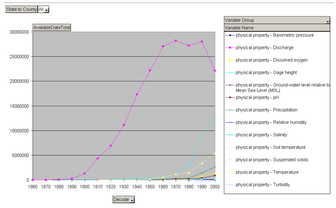
To the left are “maps” of data availability. The majority of stations is on the east coast. Stations with the longest mean period of records for a one degree cell are distributed across the country.

To the right, data availability for parameters are shown. The top show changes in phosphate/ phosphorus/ orthophosphate evolved over time. The lower shows that ‘discharge’ dropped, as ‘gauge height’ increases

Changes over time



Different types of nutrients by decade



Some physical properties by decade: Available Data Total

EPA STORET

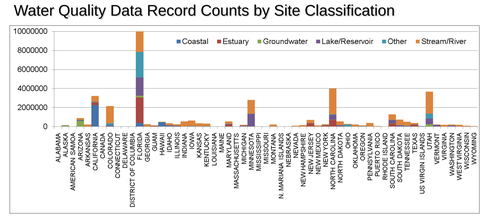
The EPA STORET datacube contains 273K sites and 2.7M series.

93% of the series are water quality data.

Organization Name	EPA
States	All

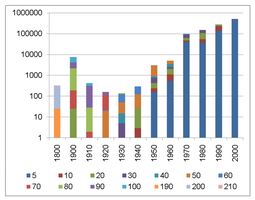
Row Labels	Series Catalog Count
Groundwater Level	285
Meteorology	34186
Oceanography	207
Other	26669
Reservoir Storage	17
Soil Quality	95955
Stage/Gage Height	6134
Streamflow	19854
Water Quality	2598880
Grand Total	2782187

Water Quality Data Record Counts by Site Classification



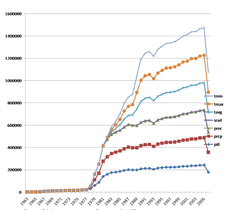
Florida is the source of about 25% of the total records.

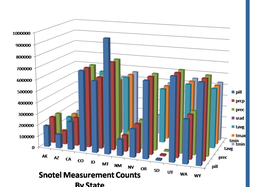
About 60% of the water quality records are short term measurements (one year or less in duration). The starting decade of the longer series are above.



Number of years of record by start decade

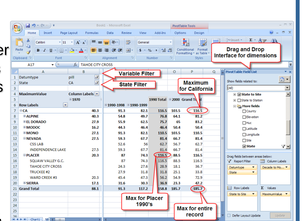
USDA SNOTEL





The SNOTEL datacube contains 810 sites, 4552 series, and 28 million observations. The number of observations increases over time, and that data is collected across a standard set of variables.

The SNOTEL datacube also has values, so you can rapidly investigate aggregations over county, states, and the entire dataset.



Conclusion

OLAP datacubes allow for domain scientists to access data simply with familiar desktop tools such as Excel PivotTables. Neither the scientist nor the programmer need write all the queries necessary to mine the data aggregates across a number of dimensions.

We loaded several Observation Database Model catalogs, basically, site summary information (often called, period of record, or series) into OLAP cubes. We conducted spatial visualization by binning into 1 degree cubes. This visualization is useful to understand the nature of the dataset. Future work will be to construct domain specific hierarchies such as watershed determined by HUC codes.

We focus here on datacubes for catalogs. Including the data values as well as can automatically produce what are now thought of a unique data products such as Daily and Yearly statistics (Mean, Minimum, Maximum, Variance) across one or more sites filtered by quality or other attribute.

Links

CUAHSI HIS: http://www.cuahsi.org/his/
HIS Wiki @ SDSC : http://river.sdsc.edu/wiki
BWC OLAP user manual: http://bwc.berkeley.edu/UserManual/UserManual.htm
Building a Cube from a CUAHSI ODM database:
http://research.microsoft.com/~vaningen/Hydrology/BearRiver/default.htm